Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 27 – 31

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling Optimization and Computing (ICMOC-2012)

A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data

Barnali Sahu[a*], Debahuti Mishra[b]

[a]*Trident Academy of Technology, Bhubaneswar, Orissa, India*
[b]*Siksha O Anusandhan University, Bhubaneswar, Orissa, India*

Abstract

Microarray data are often extremely asymmetric in dimensionality, highly redundant and noisy. Most genes are believed to be uninformative with respect to studied classes. This paper proposed a novel feature selection approach for the classification of high dimensional cancer microarray data, which used filtering technique such as signal-to-noise ratio (SNR) score and optimization technique as Particle swarm Optimization (PSO). The proposed method is divided in to two stages. In the first stage the data set is clustered using *k*-means clustering, SNR score is used to rank each gene in every cluster. The top scored genes from each cluster is gathered and a new feature subset is generated. In the second stage the new feature subset is used as input to the PSO and optimized feature subset is being produced. Support vector machine (SVM), *k*-nearest neighbor (*k*-NN) and Probabilistic Neural Network (PNN) are used as evaluators and leave one out cross validation approach is used for validation. We have compared both of our approach and approaches using PSO in the literature. It has been demonstrated that our approach using PSO gives better result than others.

Keywords: Microarray ; Signal-to-noise-ratio; Particle swarm optimization; Support vector machine; *k*-nearest neighbour; Probability Neural Network;

1. Introduction

The DNA microarray is a way to measure the expression level of thousands of genes at the same time in a cell mixture [1]. The advent of microarray technology has provided the ability to measure the expression levels of thousands of genes simultaneously in a single experiment and made it possible that providing diagnosis for disease, in molecular level [2]. However, classification based on microarray data is very different from previous classification problems in that the number of genes greatly exceeds the number of samples, which result in the known problem of 'curse of dimensionality' and over-fitting of the training data [3]. The classification of gene expression data samples involves feature selection and classifier design. Several methods have been used to perform feature selection on the training and testing data. The two broad categories of feature subset selection have been proposed: filter and wrapper [4-5]. Although wrapper approach is more effective than filter approach, in our work we have utilized the advantage of filter approach such as gene ranking. In this paper, we are interested in *gene selection* and

* Corresponding author. Tel.: +91-8895278059; fax: +91-764-2351880.
E-mail address: sahu.barnali08@gmail.com

1877-7058 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.06.005

classification of DNA microarray data and give a comparison between two of our approaches for feature selection. To resolve redundancy in gene expression values we have used one approach i.e. sample based clustering by using *k*-means clustering algorithm and the genes (features) are being grouped into number of clusters. After clustering SNR ranking is being used to rank each gene in every cluster. The gene subset selected by taking the top scored gene from each cluster will be taken as the initial search space to find the optimized subset by applying PSO and the optimized subset is used to train different classifier such as SVM, *k*-NN and PNN. In the above process the feature subset collected without using PSO will act as training set for the above classifier and the accuracy is compared with the previous approach. Leave one out cross validation is used to validate the classifier.

2. Related Work

In [6], a novel marker gene selection approach was proposed. Firstly, some top-ranked informative genes were selected by signal-noise ratio estimation method. Then a novel discrete PSO algorithm was applied to select a few marker genes and SVM was used as evaluator for getting better prediction performance on Colon tumor dataset. The authors [7] have proposed a swarm intelligence feature selection algorithm based on the initialization and update of only a subset of particles in the swarm. In their study they had tested the algorithm in 11 microarray datasets for brain, leukemia, lung, prostate, and others. They have observed that the proposed algorithm successfully increase the classification accuracy and decrease the number of selected features compared to other swarm intelligence methods. In [8] the authors have compared the use of a PSO and a Genetic Algorithm (GA) (both augmented with SVM) for the classification of high dimensional microarray data. Both algorithms are used for finding small samples of informative genes amongst thousands of them. A SVM classifier with *10- fold cross-validation* was applied in order to validate and evaluate the provided solutions

2. Methods Used

2.1. *Signal-to-noise Ratio Score*

The SNR score identifies the expression patterns with a maximal difference in mean expression between two groups and minimal variation of expression within each group [9]. In this method genes are first ranked according to their expression levels using SNR test Statistic. The SNR is defined as follows:

Signal to noise ratio= $(\mu_1 + \mu_2) / (\sigma_1 + \sigma_2)$

Here μ_1 and μ_2 denote the mean expression values for the sample class 1 and class 2 respectively. σ_1 and σ_2 are the standard deviations for the samples in each class .

2.2. *Classification Techniques Revisited*

- *k*-Nearest Neighbour Method: The *k*-nearest neighbour (*k*-NN) method was first introduced by Fix and Hodges in 1951, and is one of the most popular nonparametric methods [10]. The *k*-nearest neighbour method consists of a supervised learning algorithm where the result of a new instance query is classified based on the majority of the *k*-nearest neighbour categories. Computation cost is quite high because distances from each query instance to all training samples need to be computed. Some indexing may reduce this computational cost.

- SVM: A technique derived from statistical learning theory is used to classify points by assigning them to one of two disjoint half spaces [12]. SVM is widely used in the domain of cancer studies, protein identification and especially in Microarray data [11]. In addition, SVM can treat data with a large

number of genes, but it has been shown that its performance is increased by reducing the number of genes [13].

- PNN: PNN classifier is an implementation of a statistical algorithm known as kernel discriminate analysis where operations are organized into a multilayered feed forward network. Advantage of PNN is only one epoch of training is needed where as the drawback is it takes lots of memory for storing the training samples hence computation of recall process slows down[14].

2.3. *Proposed Algorithm using PSO*

1. Microarray data set with, Dom(C) = {I, U}, C is the random variable for class label. I= [$\boldsymbol{x_{ij}}$]; *i represents genes and* j∈ *(1, M) samples*. U = [$\boldsymbol{x_{ij}}$]; *i represents genes and j* ∈ (1, N) samples.
2. Each gene i of the data set is clustered using k-means algorithm, where each i is associated with a cluster number.
3. For each $i \in S_1^n$, SNR (i) = $\left|\frac{\mu_I-\mu_U}{\sigma_I+\sigma_U}\right|$ is calculated. Where S represents clusters from 1 to n, n is total number of clusters.
4. Top scored SNR (i) is collected from each cluster and the significant features will act as an input to PSO algorithm

Input: Microarray gene expression data containing genes having high SNR ratio score in each cluster, samples and class labels.

Output: Optimized feature genes

Initialize population

While (number of generations, or the stopping criterion is not met)

For p = 1 to number of particles

If the fitness of X_p is greater than the fitness of *Pbest*

Then Update *pbest* = X_p

For k ∈Neighbourhood of X_p

If the fitness of X_k is greater than that of *gbest* then

Update *gbest* = X_k

Next k

For each dimension d

$v_{pd}^{new} = wv_{pd}^{old} + c_1 rand_1\left(pbest_{pd} - x_{pd}^{old}\right) + c_2 rand_2(gbest_d - x_{pd}^{old})$

If $v_{pd}^{new} \notin (V_{min}, V_{max})$ then

$v_{pd}^{new} = \max\left(\min\left(V_{max}, v_{pd}^{new}\right), V_{min}\right)$

$S(v_{pd}^{new}) = \frac{1}{1+e^{-v_{pd}^{new}}}$

If (rand<S (v_{pd}^{new})) then $x_{pd}^{new} = 1$; else $x_{pd}^{new} = 0$

$x_{pd=}x_{pd}+v_{pd}$

Next d

Next p

Next generation until stopping criterion

3. Proposed Model

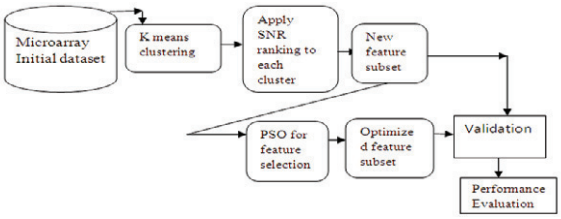

Fig. 1. Proposed model for feature selection

4. Experimental Evaluation

In our approach we have used four microarray data such as Leukaemia, Colon, DLBCL and Breast Cancer [15] to evaluate the accuracy of proposed method. The description of the above microarray datasets are given in the table1. The experiment is done in MATLAB version 7.6.0.324 (R2008a), windows XP, PC of Intel Pentium dual CPU. The data format shown in table 2 includes the dataset names, number of samples, genes, number of clusters, genes selected after *k*-means clustering and SNR ranking, and best genes selected after implementing PSO. The scatter plot for different data sets is given after implementing the proposed algorithm for feature selection and optimized feature genes are selected. Based on the optimized feature subset the various classification accuracies are listed in table 3 for different datasets. Table 4 shows the classification accuracy of different classifier for the feature genes selected without using PSO algorithm for different data sets.

Table 1. Dataset Description

Data set	**No. of genes**	**No. of instances**	**Classes**
Leukaemia	7,129	72	2
Colon	2,000	62	2
DLBCL	6,817	77	2
Breast Cancer	24,481	97	2

Table 2. Format of gene expression data

Data set	Number of samples	Genes	Number of clusters	Genes selected after *k*-means clustering and SNR ranking	Gene selected after *k*-means +SNR+PSO
Leukemia	72	7,129	500	500	10
Colon	62	2000	100	100	5
DLBCL	77	6,817	300	300	13
Breast Cancer	97	24,481	1500	1500	20

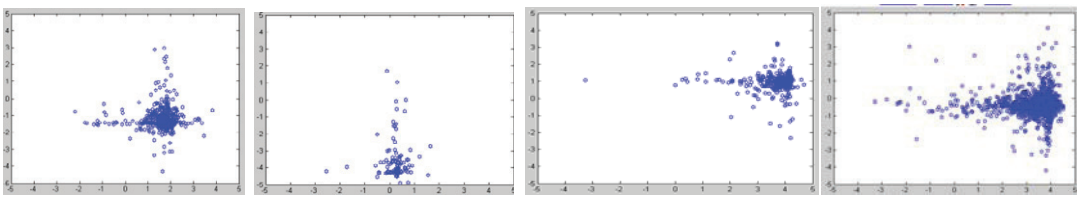

Fig. 2. (a) Scatter plot for Leukemia with n=500; (b) Scatter plot for Colon with n=100; (c) Scatter plot for DLBCL with n=300; (d) Scatter plot for Breast Cancer with n=1500

Table 3.Classification accuracy of classifiers for different datasets using PSO

Data set	Our method with PSO		
	KNN	SVM	PNN
Leukemia	99.10	100	96.34
Colon	99.44	97.50	96.00
DLBCL	99.05	100	87.32
Breast Cancer	100	100	98.17

Table 4. Classification accuracy of classifiers for different datasets without using PSO

Data set	Our method without PSO		
	KNN	SVM	PNN
Leukemia	74.87	85.49	87.18
Colon	87.01	80.19	72.23
DLBCL	61.41	84.31	88.07
Breast Cancer	83.89	2.78	78.47

5. Conclusion

In this paper we have proposed a novel approach for feature selection using PSO algorithm. Experimental results on Leukemia using SVM, DLBCL using SVM and Breast cancer using KNN and SVM have shown that our proposed algorithm is more effective and gives best result than the other. Therefore our algorithm is a useful tool for selecting feature subset for cancer microarray data.

References

[1] Y. Wang, F. S. Makedon, J. C. Ford, J. Pearlman. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics* 2005; 15:21:8: 1530-1537.

[2] T. R. Golub, D.K.Slonim, P.Tamayo, C.Huard, M. Gaasenbeek, J.P.Mesirov, H.Coller, M.L.Loh, J.R. Downing, M.A.Caligiuri,C.D.Bloomfield, and E.S.Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* 1999; 286: 531-537.

[3] E.R.Dougherty. Small sample issues for microarray-based classification, *Comparative and Functional Genomics* 2001; 2: 28-34

[4] AlanWee-Chung Liew, HongYan, MengsuYang, Pattern recognition techniques for the emerging field of bioinformatics: A review, in *Proc Pattern Recognition,* 2005; 38: 2055 – 2073.

[5] Ian B Jeffery, Desmond G Higgins, and Aedín C Culhane, Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *Proc BMC Bioinformatics* 2006: 7: 359: 1471 – 2105

[6] Hualong Yu, Guochang Gu, Haibo Liu, Jing Shen, Changming Zhu. A Novel Discrete Particle Swarm Optimization Algorithm for Microarray Data-based Tumor Marker Gene Selection. *International Conference on Computer Science and Software Engineering* 2008: pp-1057-1060

[7] Emmanuel Martinez, Mario Moises Alvarez, Victor Trevino, Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm, *Computational Biology and Chemistry* 2010: 34: 244–250

[8] Enrique Alba, Jose Garcıa-Nieto, Laetitia Jourdan and El-Ghazali Talbi. Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms, *Congress on Evolutionary Computation* 2008: version1-3.

[9] Miroslava Cuperlovic-Cuf, Nabil Belacel, Rodney. j. Ouellette, Determination of tumour marker genes from gene expression data, DDT. 2005; 10: 6: 429-437

[10] Fix, E., Hodges, J.L. Discriminatory Analysis—Nonparametric Discrimination: Consistency Properties. *Technical Report*, 21-49-004, Report no. 4, US Air Force School of Aviation Medicine, Randolph Field, 261–279.

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning*, 2002; 46: 1-3: 389–422.

[12] C. Cortes and V. Vapnik, Support vector networks. *Machine Learning*, 1995; 20: 3: 273–297.

[13] T. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machines classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 2000; 16:10: 906–914.

[14] Shamsul Huda, John Yearwood, Andrew Stranieri, Hybrid wrapper-filter approach for input feature selection using Maximum Revalance and Artificial Neural Network Input Gain Measurement Approximation. *Fourth International conference on Network and system security*, 2010:442-449.

[15] http://sdmc.lit.org.sg/GEDatasets/